A refined sampling procedure for genealogical control

Bickel, Balthasar

DOI: https://doi.org/10.1524/stuf.2008.0022

BALTHASAR BICKEL (Leipzig)

A refined sampling procedure for genealogical control

Typological distributions are the combined result of universal structural principles, areal diffusion, and shared descent. The core concern of quantitative typology is to disentangle and to identify these various factors. While areal and structural factors can be tested against each other in standard multivariate designs based on sample stratification, genealogical factors cannot be handled by sample stratification since about one third of all proven families (the strata needed) are isolates, i.e. count only one member. In response, typologists have since long sought to control for genealogical relations during sampling rather than during statistical testing. But available methods suffer from a number of drawbacks. Most importantly, they are not sensitive to the fact that different typological variable have different degrees of stability (genealogical dependence) within families, and that this again varies from family to family. This article proposes a refined method for genealogical control during sampling, which is based on DRYER's (1989) proposals but is sensitive to actual distributions within genealogical units at each taxonomic level.

1. Introduction

When exploring universal or areal skewings in the distribution of linguistic variables (also known as 'features', 'parameters', or 'characters'), a major concern is to identify and to control for inflationary effects from genealogical relatedness. If we want to argue, for example, that VO order is particularly frequent in a certain area (say, Southeast Asia) or under a certain structural condition (say, low synthesis), then we want to be able to ascertain that this distribution is due to the geographical area or to the structural condition of interest, and not due to the fact that most of the relevant VO languages happen to be members of the same genealogical family, and that this family happens to have many more members than others in the sample. Because if that were the case, the frequency of VO order (in the geographical area or structural condition under investigation could) could just as well be the result of genealogical inheritance within a few large families, rather than the result of the areal or structural factors under study. And if this so, rather than pursuing universal hypotheses about which structural conditions favor each other or about the impacts of areal diffusion, one is better advised to study pathways of genealogical inheritance in order to understand typological distributions. In short, being able to distinguish genealogical from areal and structural factors, is a fundamental issue for typology.

This problem is well-known in typology, and it has become particularly acute since the publication of the *World Atlas of Language Structures* (henceforth WALS, HASPELMATH *et al.* 2005): because the languages in WALS have been selected with many different purposes in mind (COMRIE *et al.* 2005), some families are better represented than others. Also, WALS databases containing more languages than the number of known families (e.g., those contributed by IAN MADDIESON and by MATTHEW DRYER) necessarily represent some families (e.g. Niger-Congo or Austronesian) with many more datapoints than others. Therefore, any

summary statistic of WALS raises the issue to what degree it is influenced by the size and nature of families in the database, and before interpreting the data in areal or universal terms, we obviously need to control for confounding factors from genealogy.

In this article, I first review available methods for genealogical control. Expanding on Dryer's (1989) proposals, I then develop a general algorithm that allows drawing samples of languages in which the distribution of a given typological variable can be assumed to be unaffected by direct genealogical inheritance — at least to a reasonable degree of confidence.

2. Available methods

The need to identify and control for inflationary effects from confounding factors is by no means unique to typology. The textbook solution in other disciplines is sample stratification: the dataset is divided into genealogical families (the strata) and within each of these, one randomly chooses the same number of languages. However, this not a practical approach in typology because it can lead to inflationary effects from areal variables: for example, when random sampling happens to pick Romansh as the sole representative of Romance and High Alemannic as the sole representative of Germanic, this will overestimate areal effects in Europe because the two languages are under much more intense contact with each other than, say, Portuguese and Swedish. In response to this, one might chose to admit several languages from each stratum in the hope of reducing such effects. However, this option is severely limited because about a third of the proven stocks in the world are isolates. Since strata need to contain the same number of languages, the inclusion of isolates implies that only one datapoint can be admitted for each stratum, even for non-isolates like Romance or Germanic (cf. JANSSEN *et al.* in press).

The alternative approach that has become standard in typology is to control for genealogical factors by what is called probability sampling in RIJKHOFF & BAKKER (1998) or, with a narrower focus, genealogical or genealogically-balanced sampling.[1] The basic idea is that the sample does not consist of the typological values of individual languages but of values that are representative for those genealogical units which are old and historically diverse enough so as to be reasonably independent with regard to the typology under study. For example, word order is relatively consistent within the branches of Indo-European, and it is reasonable to suspect that this distribution is at least to some degree caused by common descent within each branch. Therefore, including more than one datapoint per branch (e.g. Balto-Slavic, or Indo-Iranian) into a sample would create an inflationary effect

[1] The alternative is variety sampling, which does not serve to control genealogical factors, but aims at taking a snapshot of current areal and genealogical distributions. For such undertakings, one will want to have languages represented in proportion to family and area sizes—resulting in the exact opposite of genealogically-balanced samples. As BAKKER & RIJKHOFF (1998) note, variety sampling is not suitable for hypothesis testing but has its merits in exploratory qualitative research.

from genealogy, and so, under genealogical sampling, each branch is represented by one datapoint only. The same danger of an inflationary effect does not arise to the same extent on higher taxonomic levels: word order *between* branches varies strongly in Indo-European, and there is considerably less confidence that each particular order derives directly from Proto-Indo-European without any areal or structural factor favoring or disfavoring particular developments. In order to find out about these factors, it is therefore reasonable to include several datapoints from the entire stock, viz. one per branch (or perhaps some higher node in the tree which shows word order uniformity).

Under genealogical sampling, then, one would admit one datapoint per branch. Depending on one's theory of genealogical relationships and historical stability of typological variables, instead of the branch, the genus (in DRYER's 1989 sense), family, or stock level may just as well be used as the sampling unit, or one might want to vary the sampling level depending on the age and taxonomic diversity of each stock (see RIJKHOFF & BAKKER 1998 for some suggestions on this). In any event, the value of a datapoint is sought to be representative of the relevant sampling unit (the genus, the stock, etc.). Representativeness is achieved by selecting what is found to be the modal, or most archaic, or in another way typical language. (More often than not this in practice based on impressionistic assumptions about the behavior of the relevant unit with regard to the variable under consideration.) The benefit of this procedure is that one can construct an all-purpose sample before data collection and then build a database by simply working through all languages in the sample, coding for any variable of interest.

A problem with this approach of all-purpose sampling has been noted *en passant* by DRYER (1989): if we find that one genealogical unit, e.g. a genus, happens to be diverse in the variable of interest (say, there are both language with OV and VO word order in the genus), this by itself suggests that the distribution within that genus may not, or not completely, depend on genealogical relatedness, after all. In such a case, there is some justification for allowing more than one datapoint to the sample from this genus. DRYER proposes to admit all *distinct* datapoints. For example, if a genus contains both OV and VO languages, one would admit both an OV datapoint and a VO datapoint. As a result of this, the sampling units are no longer identical with genealogical units (let alone their proto-languages, cf. DRYER 2000) and can vary from genus to genus. If a genus is diverse, it will provide many sampling units (datapoints), but if it is homogenous, it will provide only one unit. And if the genus contains only one member, i.e. an isolate, that language will provide one unit. All that matters is that the sampling units can reasonably well be seen as genealogically independent with respect to the variable of interest, i.e. as not sharing values because of shared inheritance.

This procedure, which I call here controlled genealogical sampling, has a number of advantages over both all-purpose genealogical sampling and stratified random sampling. First, controlled genealogical sampling eschews the thorny question of what is the best representative of a branch or genus (or even stock, if that was cho-

sen as the sampling unit). Second, under controlled genealogical sampling, the question of how much a typological value depends on shared descent can now be answered separately for each variable in each family. This fits well with the finding that different variables have different degrees of historical stability in families (Nichols 2003, among others), and also that the same variable may be more stable in one family than in another: if a variable is diverse in Sino-Tibetan but not Indo-European, then the distribution of that variable is less likely to exclusively depend on the proto-language in Sino-Tibetan than in Indo-European, and so Sino-Tibetan will provide more datapoints. With all-purpose sampling the same degree of stability needs to be assumed for all typological variables in all families, and the estimate of that degree fixes the taxonomic level at which sampling is performed (typically the major branch or genus level). Third, unlike with stratified random sampling, isolates do not pose a problem: they simply provide an independent unit, and this has no consequences on how much non-isolates can contribute to the sample. Under controlled genealogical sampling, sample size is entirely a matter of how likely it is that a typological distribution depends on genealogy: the more we suspect such a dependency for a given variable, the smaller the sample needs to be. In turn, if there is no reason to suspect such a dependency, the sample can be as rich as the dataset.

The drawback of controlled genealogical sampling is that survey work cannot be limited to a preselected sample but must look in detail at within-branch (or at least within-stock) variance. However, for many typological research questions it has become crucial to study within-family variance anyway. This is essential for example, if one wants to estimate historical stability and transition probabilities, or if one wants to determine the contribution of genealogical factors on typological distributions (see MASLOVA 2000, NICHOLS 2003, BICKEL & HILDEBRANDT 2005, and below, Section 4, for some recent research addressing such issues). And, since many WALS and other databases are not pre-sampled genealogically, it has become possible to investigate within-genealogy variance and create *post-hoc* samples on the basis of this.

However, there are a number of open problems in the procedure of how to choose languages in such a situation, and in the following I propose solutions to them. Along the lines of this solution, I formulate a general algorithm (implemented in open-source software) that allows one to perform genealogical *post-hoc* sampling on any sufficiently rich database, with any kind of genealogical taxonomy.

2. Problems in controlled genealogical sampling

2.1 Non-discreteness

In classical controlled genealogical sampling according to DRYER (1989), the diversity within genera is an all-or-nothing issue. If there is only one value of the

variable in the genus (or any other genealogical unit, for that matter), i.e. there is no diversity, the genus provides only one sampling unit. If a genus shows more than one value of the variable of interest, i.e. there is some diversity, then all attested values are taken as sampling units, regardless of their actual distribution within the genus. But there are two kinds of diversity: (a) chance diversity, where values have a uniform (e.g. 50% : 50%) distribution in the unit, and (b) statistically significant skewings, where one value dominates (e.g. 90% : 10%). The distinction can easily be assessed with a statistical test at a chosen significance level, and it affords distinct treatment in sampling.

If we find scenario (a), it is likely that genealogical membership is irrelevant for the distribution of values in the unit, and there is no reason to include only one datapoint per distinct value in the sample. Instead, each language can be included without any reasonable risk of a genealogical inflation effect. This is desirable because it means that one loses less of the information in the dataset than would be the case when following DRYER's original method, where the number of sample units is equal to the number of distinct types (regardless of token frequencies). Losing less information in turn allows a better assessment of those areal or structural factors that are hypothesized to drive the observed distribution.

If we find a significant skewing as in (b), by contrast, it is likely (though by no means necessary!) that the distribution is induced by shared retention or innovation—i.e. it is a skewing that what we want to control for.[2] This case should be treated the same way as a totally homogenous genealogical unit, and so the majority value should be included in the sample only once for the unit. Note, however, that this does not apply to the minority value, even if it occurs several times: the presence of the minority value must be due to some non-genealogical factor, i.e. perhaps it was precisely an areal or structural factor under investigation that triggered the deviation. To find out, each such deviating language needs to be included in the sample. Of course, if the deviations themselves make up a genealogical subgroup within the genus, the deviating pattern is a shared innovation, and we again need to control for an inflationary effect from that subgroup by admitting only one datapoint to the sample. This is a more general issue and the topic of the following section.

2.2 Genealogical levels

In its classical form, controlled genealogical sampling operates on a single predetermined taxonomic level, e.g. the genus or the stock. However, as just noted, when we find genealogical skewing at one level, e.g. the genus (e.g. dominance of

[2] Note that distributional skewing is used here as a criterion for whether or not it is reasonable to control for a genealogical confounding factor in a sample. Distributional skewing is not a criterion to decide whether something should be reconstructed. In other words, sampling under genealogical control is methodologically very different from genealogical reconstruction and it has completely different goals.

OV over VO), the minority pattern (VO) might be genealogically skewed at the next lower level (the sub-genus), or it might again be distributed independently of genealogical relations. Also, when we find genuine (non-skewed) diversity (e.g. half VO, half OV) in a genealogical unit, we cannot be sure that the distribution of values is not skewed by a taxonomic level just below the one we looked at, i.e. that perhaps there are two subgroups, one with 90% VO and one with 10% VO. Alternatively, the two values might distribute equally across the branches, with no detectable skewing. Table 1 illustrates this by an imaginary genus with 10 members, half of which are VO and half OV.

Table 1. Two possible distributions in a heterogeneous genus.

		OV languages	VO languages
	total genus	10	10
Scenario (a)	subgenus 1	9	1
	subgenus 2	1	9
Scenario (b)	subgenus 1	5	5
	subgenus 2	5	5

Since the goal is to control for any genealogical effect in the sample, we also need to control for such possible subgenus effects, and, recursively, for sub-sub-genus etc. (until we reach the lowest taxonomic level, i.e. a language or dialect). Therefore, in scenario (a) of Table 1, one would want to include each subgenus as a sampling unit (i.e. two sampling units for this genus), but no more, because we find that the distribution of values is significantly affected by subgenus membership. But in scenario (b), the distribution is unlikely to be genealogically induced (we cannot predict the values from subgenus membership), and so one can include all languages of the genus without risking a genealogically-induced inflation effect. If one were to treat scenario (b) like (a), and therefore would include only two datapoints (one OV, one VO), one would lose data that is perhaps relevant for detecting areal or universal skewings. In fact, areal skewings would be most likely precisely in a scenario like (b), e.g. when both subgenera straddle two areas, one VO and the other OV.[3] Such an effect will be more difficult to detect in a reduced sample of two datapoints than in the full set of 20 datapoints.

Going up the taxonomy, i.e. from genus to stock and phylum, a similar issue arises: genealogical skewings can be induced just as well by higher levels as by lower levels. If for some variable of interest all genera of a stock have the same values, this is perhaps because they are related and the value is inherited. Again, this can be controlled for by looking at the distribution of the variable and then treat the entire stock as one sampling unit when there is a significant skewing (and therefore a possibility of a genealogical inheritance effect), but not when the distribution is even.

[3] In reality, for many variables we in fact expect that genealogical skewings become stronger the lower the taxonomic level is. Word order, for example, is more likely to reflect inheritance over a few hundred than over a few thousand years. Scenario (a) is thus more likely to occur than scenario (b).

3. An improved algorithm for controlled genealogical sampling

The solution to the discreteness problem (Section 2.1) can be solved by replacing the all-or-nothing diversity criterion by one that tests for statistical skewing. The level problem (Section 2.2) can be solved by replacing single-level sampling by a recursive sampling procedure that works through all levels of the genealogical taxonomy. Thus, we can develop a general algorithm that draws a genealogically controlled sample by recursive testing for diversity at each taxonomic level and admit genealogical units (from stock to language) to the sample only if they are significantly distinct from their sister units at each level with regard to the typological variables of interest. A bonus effect of such an algorithm is that it also estimates the genealogical stability of the relevant typological variables: when the algorithm finds many skewing effects at a given taxonomic level, this suggests relativily strong historical stability at that level; if there are few, this suggests relatively less stability. (Such estimates of stability are a useful tool to assess distributional hypotheses, as we will see in Section 4.)

In the following I describe an algorithm that implements this idea. The alogorithm creates a controlled genealogical sample for a given typological response variable (e.g. position of relative clauses) on which we wish to test a set of predictor variables (e.g. a structural variable such as verb-object order and an areal control variable such as new-world vs. old-world languages). The algorithm is general, i.e. it can be applied to any variable and any genealogical taxonomy. All that is needed is a table, such as the ones contained in WALS, where each language is coded for each taxonomic level one wants to control for, and, of course, for the typological variables of interest. (If the variables have missing typological values for a language, that language will simply be ignored by the algorithm.) The algorithm produces a controlled genealogical sample ('g-sample') consisting of what I call 'g-units' (short for genealogical units), defined as those genealogical units that have typological values unlikely (e.g. at a 5% level of statistical significance) to be induced by shared descent from the next higher genealogical unit. The core of the algorithm is the following procedure.[4]

> We begin with a genealogical unit U at a chosen taxonomic level t (U_t) (t_0 = highest taxon, t_n = language or dialect), coded for a typological response variable V. In a first step (referred to as Routine A below), we determine $N(U_t)$, the number of units in U_t. There are two cases: (i) if $N(U_t) = 1$, i.e. there is only one language in U_t, we simply add a g-unit to the g-sample, with the values on V and P associated with that language. (ii) If $N(U_t) > 1$, we test (Routine B) whether the members of U_t are all the same with regard to V or at least deviate significantly from expectations under independence within V, i.e. whether a skewing effect on V from genealogical relatedness in U_t is likely. A suitable significance test is a randomized one-sample χ^2-test for given probabilities (as described and

[4] This algorithm is implemented in the open-source package *R* (R DEVELOPMENT CORE TEAM 2005). The source code is available from my webpage at <http://www.uni-leipzig.de/~bickel>.

motivated in JANSSEN *et al.* in press).[5] If the test is negative, it is unlikely (to a reasonable degree) that U_t has undergone a skewing effect from genealogical relatedness. In that case, we start over with Routine A at the next lower taxonomic level U_{t+1} within U_t, until we reach a level t for which $N(U_{t+1}) = 1$, or the test is positive. If the test is positive, the members of U_t all have the same values or the value distribution deviates strongly from expectations under independence. Therefore, it is reasonable to suspect a skewing effect caused by genealogical relatedness, and we add only one g-unit with the majority (modal) value on V (i.e. the most frequently found value of V in U_t) to the g-sample, together with the associated values on P. If there is more than one level of P (e.g. languages from two areas, if areas is the predictor) associated with the majority value of V in U_t, add one g-unit for each level of P, but all with the same (i.e. majority) value on V: this ensures that counterevidence against effects of P on V is not lost. We next look at the distribution of the minority ('deviating') values on V in U_t at the next lower taxonomic level: we test in each such unit U_{t+1} whether it is dominated by any of the deviating values from U_t, and if so, whether this dominance is statistically ignificant. If that is the case, it is likely that what appears as a deviating minority pattern at level U_t has a shared history at U_{t+1}, and U_{t+1} should only contribute one g-unit with the deviating value (or at most as many as the deviating value combines with different predictor levels). If, by contrast, unit U_{t+1} is not dominated by any of the deviating values from U_t, we move to the next lower level in order to find out whether there is any subgroup in which the deviating pattern appears as the result of shared descent until we again reach a level $t+1$ for which $N(U_{t+1}) = 1$. Once we reach this level, we move the next unit U_t and start over with Routine A, until all U_t's have been worked through.

In order to control for all reasonably well-known genealogical relationships, one would typically want to start with $U_t = U_0$, the highest accepted taxon (the stock or phylum, based on what taxonomy one works with). If one thinks that genealogical relationships older than the genus are *a priori* unlikely to have an effect on the typological distribution at hand, then one might want to start with t = genus, following DRYER (1989).

4. An example

Humans, linguists among them, are highly gifted in detecting visual patterns. When browsing WALS and, even more so, when exploring its interactive version, one is bound to see all sorts of areal and macro-areal patterns. Many of them are flukes because there is no theory of historical genesis behind them, and no clear hypothesis to test (Bickel & Nichols 2005a, 2006), but many others follow from

[5] I assume that for most practical purposes, scalar variables will have only very few values per g-unit so that they are best analysed as multinomial types and subject to a χ^2-test. When this is not the case, the χ^2-test can be replaced by a test comparing the within-U_t variance to the variance expected from the total distribution or from theoretical grounds.

received theories and afford straightforward statistical testing. One such theory is the Eurasian areality theory, first proposed by Jakobson (1931) and explored in ongoing work by Bickel and Nichols (e.g. Bickel & Nichols 2003, 2005a, 2005b, 2005c; Nichols & Bickel 2005). The Eurasian area combines all of the large spread zones (in Nichols' 1992 sense) in the north, south and southeast of Eurasia and is characterized by a relatively 'flat' typological profile that contrasts with the rich structural diversity of Africa, the Americas, the Pacific, and 'enclave' regions in the Caucasus and the Himalayas. For the geographical definition and the proposed genesis of the Eurasian area, see Bickel & Nichols (2003, 2005c).

Dryer's (2005) WALS data on negation markers suggests that one specific type, double negation, follows this pattern, too: both the visual impression (Map 1) and the raw numbers (Table 2) from the atlas suggest an extremely depressed proportion of double-negation languages in Eurasia that contrasts with the rest of the world.

Map 1. Languages with double (black dots) vs. simple (white dots) negation according to Dryer (2005)

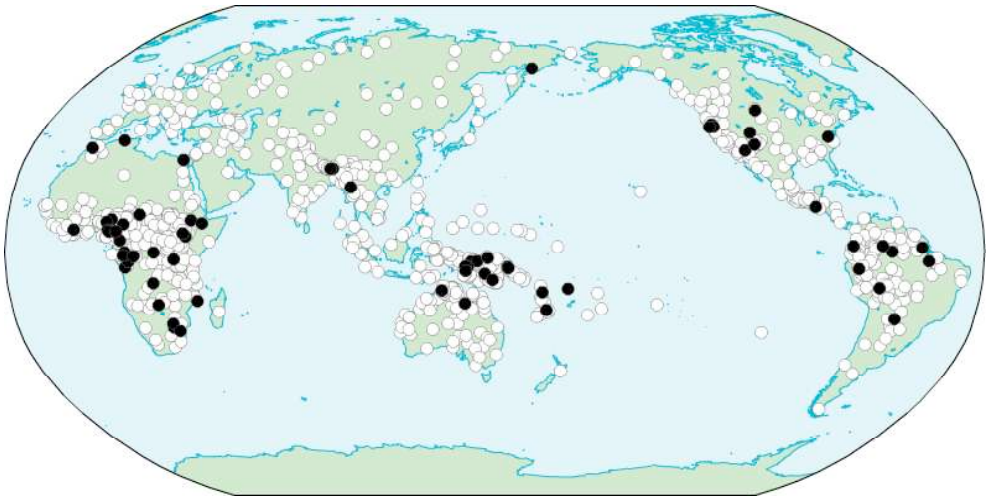

Table 2. The distribution of negation types over Eurasia vs. the rest of the world (Unsampled dataset from Dryer 2005)

	simple NEG	double NEG	Total
Within Eurasia	222 (99%)	3 (1%)	225
Outside Eurasia	723 (92%)	63 (8%)	786
Total	945	66	1011

The distribution in Table 2 is statistically highly significant under a Fisher Exact Test (p = .0001). However, the largest families in Eurasia contribute large numbers of identical datapoints to the sample: Indo-European contributes 53, Altaic (recognized as a family in WALS) 23 and Uralic 13 datapoints with simple negation; Sino-Tibetan adds 66 languages with simple and 3 (Limbu, Lepcha and Burmese) with double negation. This raises the possibility that the 99% percent dominance of simple-negation languages in Eurasia is due to the fact most languages in the Eurasian dataset inherited simple negation from their respective proto-languages. If so, there is no evidence for an areal factor to have played a direct role. If, by counter-hypothesis, the skewing in Table 1 is not due to genealogical relations, it should be replicable in a genealogically controlled sample.

To test this counter-hypothesis, I applied the algorithm proposed in Section 3 to Dryer's dataset. The algorithm reduces families with total or near-total homogeneity to a single datapoint because their within-family distribution is the likely result of common descent. Deviating patterns are tracked to the degree that they are not themselves the result of shared descent. In the dataset at hand, no deviating pattern showed genealogical skewing at lower taxonomic level (e.g. within Bantu), but one cannot exclude the possibility that a more fine-grained taxonomy would detect such skewing. The result of the algorithm shows that most (non-isolate) families and one genus (Algonquian) show significant genealogical skewing, either statistically (8, mostly large families) or absolutely (44, mostly small families and the Algonquian genus). But 12 families and 3 genera do not show evidence for such skewing, and there is no reason therefore to limit their contribution to the g-sample (ranging from 2 datapoints in the case of Quechuan to 11 in the case Arawakan). The resulting g-sample is shown Table 3.

Table 3. The g-sampled distribution of negation types over Eurasia vs. the rest of the world

	simple NEG	double NEG	Total
Within Eurasia	14 (82%)	3 (18%)	17
Outside Eurasia	157 (71%)	63 (29%)	220
Total	171	66	237

The distribution in Table 3 is not significant statistically (p = .41) and therefore does not replicate the results based on the unsampled dataset in Table 2.

However, since the algorithm identified a substantial number of skewed families (N = 51), the questions arises whether there could not be another kind of areality effect at work: instead of directly influencing the distribution of typological variables, areal relationship can also be expected to increase the general stability of such variables, on the assumption that a family is more likely to keep a typological feature if it is in regular and long-term contact with families that have the same

feature. Thus, if the (non-singleton) families in Eurasia are significantly more often genealogically skewed than those outside, areality might have played in role in leading to increased stability. Since the g-sample algorithm determines whether families are skewed in the dataset or not, this alternative possibility can be directly tested on the output of the algorithm. The result is tabulated in Table 4, which shows no statistical evidence for the hypothesis (p = .19), i.e. is fairly close to what one can expect from the margin totals.

Table 4. Families with vs. without genealogically skewed distributions

	not skewed	skewed	Total
Within Eurasia	0 (expected 2)	10 (expected 8)	10
Outside Eurasia	12 (expected 10)	41 (expected 42)	53
Total	12	51	63

There is good reason, therefore, to suspect that the distribution suggested by the visual inspection of Map 1 is due to inflationary effects of large families in Eurasia and does not reflect the areal trend predicted by the Eurasian areality theory, under any interpretation of the possible effect of areal relations on typological distributions.

While Table 3 does not replicate the unsampled dataset, it replicates the results obtained on an all-purpose genealogical sample, which seeks to include one datapoint per stock and one per major branch if the stock is old and genealogically diverse. Such a sample, again coded for double vs. simple negation,[6] is available in AUTOTYP (the 'GEN1' sample, as used in, e.g., Bickel & Nichols 2005c). The results are shown in Table 5.

Table 5. The distribution of negation types over Eurasia vs. the rest of the world, according to AUTOTYP (Bickel & Nichols 1996ff)

	simple NEG	double NEG	Total
Within Eurasia	44 (92%)	4 (8%)	48
Outside Eurasia	131 (84%)	25 (16%)	156
Total	175	29	204

Like in Table 3, the distribution in the AUTOTYP sample does not show evidence for a statistically significant association of negation type and area (Fisher Exact Test, p = .24).

[6] There are minor coding differences between the WALS and the AUTOTYP datasets: Belhare (Sino-Tibetan), for example, is coded as having double-negation in AUTOTYP because this is the majority strategy in the paradigm, while in WALS, the same language is coded as having simple negation. Tundra Nenets (Uralic) is coded as having double-negation in AUTOTYP on account of its auxiliary plus connegative strategy, while in WALS the connegative is analyzed differently, and the language is therefore coded as having simple negation. As the test results show, these coding differences have no impact on the overall findings.

The fact that the results of the g-sample based on the 1011-languages dataset in WALS can be replicated by an all-purpose 204-languages sample validates the use of such smaller samples to some degree. For practical reasons, it is often impossible to collect the large datasets that are needed to derive g-samples, and working with an all-purpose genealogical sample is a good alternative. As Table 5 shows, such a sample allows a reasonable estimate of the over-all distribution. Still, one cannot be certain that such estimates are valid for any given typological variable, and full testing of hypotheses clearly needs larger datasets.

5. Conclusions

The applied example presented in the preceding section suggests that visual impressions and also statistical analyses of unsampled datasets can easily lead one to mistake genealogically-induced distributions for the result of areal (or structural) factors. This demonstrates the need for genealogical control in sampling (pace WIDMAN & BAKKER's (2006) recent vote for random sampling). While all-purpose genealogical samples are practical alternatives for estimating distributions, proper hypothesis-testing requires large datasets, with extensive within-family coverage, and the application of a post-hoc sampling procedure that is based on the actual distribution of datapoints within genealogical units at all known taxonomic levels.

As pointed out by JANSSEN *et al.* (in press), it is important to bear in mind that genealogical sampling, whether controlled or not, is fundamentally different from random sampling. Under random sampling procedures, representativeness is ensured by independent data selection. Under genealogical sampling, representativeness can be ensured only by aiming at technical exhaustiveness: by aiming at samples that approximate full coverage of the population (e.g. all known languages of the world). JANSSEN et al. (in press) argue that under this condition, only distribution-free methods can be applied, specifically randomization and exact tests. This is a necessary consequence of applying genealogical sampling procedures. If one wishes to avoid this, and use statistical methods based on classical sampling theory, one needs to perform random sampling. But then, as argued in the introduction, it becomes very difficult, perhaps impossible, to control for genealogical confounding factors.

Acknowledgments

I'd like to thank Dik Bakker and Michael Cysouw for helpful comments on an earlier draft and to my student assistant Taras Zakharko for implementing the algorithm in R.

References

Bickel, B. & K. Hildebrandt 2005. Diversity in phonological domains. Paper presented at the 6th Biannual Conference of the Association for Linguistic Typology, Padang, July 2005; available at http://www.uni-leipzig.de/~autotyp/download.

Bickel, B. and J. Nichols 1996ff. The AUTOTYP database. Electronic database; http://www.uni-leipzig.de/~autotyp.

Bickel, B. & J. Nichols 2003. Typological enclaves. Paper presented at the 5th Biannual Conference of the Association for Linguistic Typology, Cagliari, September 18; available at http://www.uni-leipzig.de/~autotyp/download.

Bickel, B. & J. Nichols 2005a. Areal patterns in the World Atlas of Language Structures. Paper presented at the 6th Biannual Conference of the Association for Linguistic Typology, Padang, July 24; available at http://www.uni-leipzig.de/~autotyp/download.

Bickel, B. & J. Nichols 2005b. Inflectional synthesis of the verb, in In Haspelmath, M., M. S. Dryer, D. Gil & B. Comrie (eds.) *The world atlas of language structures*, 94 – 97. Oxford: Oxford University Press.

Bickel, B. & J. Nichols 2005c. Inclusive/exclusive as person vs. number categories worldwide, in Filimonova, E. (ed.), *Clusivity,* 47 – 70, Amsterdam: Benjamins

Bickel, B. & J. Nichols 2006. Oceania, the Pacific Rim, and the theory of linguistic areas. *Proceedings of the 32nd Annual Meeting of the Berkeley Linguistics Society*.

Comrie, B., M. S. Dryer, D. Gil & M. Haspelmath 2005. Introduction. In Haspelmath, M., M. S. Dryer, D. Gil & B. Comrie (eds.) *The world atlas of language structures*, 1 - 8. Oxford: Oxford University Press.

Dryer, M. S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13, 257 – 292.

Dryer, M. S. 2000. Counting genera vs. counting languages. *Linguistic Typology* 4, 334 – 350.

Dryer, M. S. 2005. Negative morphemes. In Haspelmath, M., M. S. Dryer, D. Gil & B. Comrie (eds.) *The world atlas of language structures.* Oxford: Oxford University Press.

Jakobson, R. 1931. K xarakteristike èvrazijskogo jazykovogo sojuza, *Selected Writings 1*, The Hague: Mouton 1970, 144 – 201.

Janssen, D., B. Bickel & F. Zúñiga 2006. Randomization tests in language typology. *Linguistic Typology* 10, 419 – 440.

Maslova, E. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4, 307 – 333.

Nichols, J. 1992. *Language diversity in space and time*. Chicago: The University of Chicago Press.

Nichols, J. 2003. Diversity and stability in language. In Janda, R. D. & B. D. Joseph (eds.) *Handbook of Historical Linguistics*, 283 – 310. London: Blackwell.

Nichols, J. & B. Bickel 2005. Possessive classification, In Haspelmath, M., M. S. Dryer, D. Gil & B. Comrie (eds.) *The world atlas of language structures*, 242 – 45. Oxford: Oxford University Press.

R Development Core Team 2005. *R: a language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing (www.r-project.org).

Rijkhoff, J. & D. Bakker 1998. Language sampling. *Linguistic Typology* 2, 263 – 314.

Widmann, T. & P. Bakker 2006. Does sampling matter? A test in replicability, concerning numerals. *Linguistic Typology* 10, 83 – 95.

Correspondence address

Balthasar Bickel
Institut für Linguistik
Universität Leipzig
Beethovenstrasse 15
04015 Leipzig
www.uni-leipzig.de/~bickel
bickel@uni-leipzig.de